Odhalení skryté struktury a vnitřních vazeb dat metodami PCA, FA a CLU vícerozměrné statistické analýzy

Prof. RNDr. Milan Meloun, DrSc.,
Katedra analytické chemie, Univerzita Pardubice, 532 10 Pardubice, milan.meloun@upce.cz
a
Prof. Ing. Jiří Militký, CSc.,
Katedra textilních materiálů, Technická univerzita Liberec, 461 17 Liberec, jiri.militky@tul.cz

Souhrn: *Na řadě praktických úloh jsou ukázány diagnostické vlastnosti vybraných metod vícerozměrné statistické analýzy, především metody hlavních komponent, faktorové analýzy a tvorby shluků. Uvedenými metodami je odhalena vnitřní struktura v datech, skryté vnitřní vazby a řada souvislostí ukrytých mezi znaky a mezi objekty.*

Odhalení struktury ve znacích a objektech: Zdrojová matice dat $\boldsymbol{X}$ má rozměr $n \times m$. Před vlastní aplikací vhodné metody vícerozměrné statistické analýzy je třeba vždy provést *průzkumovou (exploratorní) analýzu dat*, která umožňuje: a) posoudit *podobnost objektů* pomocí rozptylových a symbolových grafů, b) nalézt *vybočující objekty*, resp. jejich znaky, c) stanovit, zda lze použít předpoklad *lineárních vazeb,* d) ověřit *předpoklady o datech* (normalitu, nekorelovanost, homogenitu).

Jednotlivé techniky k určení vzájemných vazeb se dále dělí podle toho, zda hledají 1. *strukturu a vazby ve znacích* nebo 2. *strukturu a vazby v objektech*. To znamená a) hledání struktury ve znacích v metrické škále – *faktorová analýza FA*, *analýza hlavních komponent PCA* a *shluková analýza*. b) hledání struktury v objektech v metrické škále – *shluková analýza*. c) hledání struktury v objektech v metrické i v nemetrické škále – vícerozměrné škálování. d) hledání struktury v objektech v nemetrické škále – *korespondenční analýza.*

Určením struktury a vzájemných vazeb mezi znaky ale i mezi objekty se zabývají techniky redukce znaků na latentní proměnné, *metoda* analýzy hlavních komponent (*PCA*), *metoda faktorové analýzy (FA)* a shlukování (*CLU*).

Maticový diagram. Diagram ukazuje rozptylové diagramy jednotlivých dvojic znaků. Je zřejmé, že vysoké hodnoty korelačního koeficientu vedou ke zřetelné lineární závislosti (přímka), zatímco nízké hodnoty ukazují, že osoby nejsou v grafu zobrazeny na přímce ale nacházejí se spíše v chaotickém mraku bodů.

Směry maximálního rozptylu v rozptylovém diagramu. Intuitivně tušíme, že by bylo nejlepší proložit objekty novou souřadnou osu tak, aby byla totožná s nejtěsněji prokládanou přímkou. Nejtěsněji objekty prokládaná přímka, a tím pádem i nová souřadnicová osa zvaná *PC*1, leží ve směru *největšího rozptylu objektů*. Souřadnicová osa *PC*1 není obecně přitom paralelní s žádnou z původních

os znaků x_1, x_2 a x_3. Novou souřadnou osu $PC1$ nazveme v následujících kapitolách *první hlavní komponentou.* Bude ležet *ve směru maximálního rozptylu* zdrojové matice dat $\boldsymbol{X}$. S touto osou je spjata jakási nová latentní proměnná, jejíž význam zatím neznáme. Metoda hlavních komponent PCA nám poskytne nejenom tuto první komponentu, ale také další hlavní komponenty a je jenom na nás, jak je budeme interpretovat, co znamenají v analýze a co popisují. Obvykle jedna nebo dvě, maximálně však tři, nové souřadnicové osy $PC1$, $PC2$ a $PC3$ vystihnou dohromady největší podíl proměnlivosti dat.

První hlavní komponenta proložením nejmenšími čtverci. Účelem je proložit přímku v prostoru x_1, x_2 a x_3 umístěnými objekty. Když z každého i-tého objektu spustíme k této přímce kolmici, obdržíme vzdálenost e_i zvanou *reziduum.* Nejtěsněji proloženou přímku pak dostaneme, když suma čtverců reziduí bude dosahovat své minimální velikosti

$$RSS = \sum_{i=1}^{n} e_i^2 = \min.$$

Existují tedy dvě kritéria ke konstrukci první hlavní komponenty: 1. Směr největšího rozptylu, 2. metoda nejmenších čtverců. Ukazuje se, že obě metody vedou ke stejnému výsledku. Může nastat případ, že první hlavní komponenta $PC1$ nestačí dostatečně popsat rozptýlení objektů. Druhá hlavní komponenta $PC2$ je ortogonální vůči první (je na ni kolmá) a je umístěna do směru druhého největšího rozptýlení objektů v rovině. Podobně, vytvoří-li objekty v prostoru přibližný elipsoid, budou do prostorových os elipsoidu umístěny tři osy vystihující největší rozptýlení objektů čili osy tří hlavních komponent $PC1$, $PC2$ a $PC3$. Nové tři osy nekorelují, jsou vzájemně ortogonální, tj. na sebe kolmé. Metoda hlavních komponent PCA umožňuje tuto geometrickou projekci zobecnit na libovolný počet m znaků.

Zaměření metody PCA: Poprvé byla zavedena Pearsonem již v roce 1901 a nezávisle Hotellingem v roce 1933. Cílem analýzy hlavních komponent je zjednodušení popisu skupiny vzájemně lineárně závislých neboli korelovaných znaků čili rozklad zdrojové matice dat do *matice strukturní* a do *matice šumové.* V analýze hlavních komponent nejsou znaky děleny na závisle a nezávisle proměnné jako v regresi. Techniku lze popsat jako metodu lineární transformace původních znaků na nové, nekorelované proměnné nazvané *hlavní komponenty.* Každá hlavní komponenta představuje lineární kombinací původních znaků. Základní charakteristikou každé hlavní komponenty je její míra variability čili *rozptyl.* Hlavní komponenty jsou seřazeny dle důležitosti, tj. dle klesajícího rozptylu, od největšího k nejmenšímu. Většina informace o variabilitě původních dat je přitom soustředěna do první komponenty a nejméně informace je obsaženo v poslední komponentě. Platí pravidlo, že má-li nějaký původní znak malý či dokonce žádný rozptyl, není schopen přispívat k rozlišení mezi objekty.

Standardním využitím PCA je snížení dimenze úlohy čili redukce počtu znaků bez velké ztráty informace, a to užitím pouze prvních několika hlavních

komponent. Toto snížení dimenze úlohy se netýká počtu původních znaků. Je výhodné především pro možnost zobrazení vícerozměrných dat. Předpokládá se, že nevyužité hlavní komponenty obsahují malé množství informace, protože jejich rozptyl je příliš malý. Tato metoda je atraktivní především z důvodu, že hlavní komponenty jsou nekorelované. Namísto vyšetřování velkého počtu původních znaků s komplexními vnitřními vazbami analyzuje uživatel pouze malý počet nekorelovaných hlavních komponent. Analýza hlavních komponent je rovněž součástí průzkumové analýzy dat. Snížení rozměrnosti je často využíváno při konstrukci komplexních ukazatelů jako lineárních kombinací původních znaků. Například první hlavní komponenta je vlastně vhodným ukazatelem jakosti, pokud původní znaky charakterizují její složky. Využití první hlavní komponenty jako komplexního ukazatele je běžné v oblasti ekonomie, sociologie a medicíny. První dvě respektive první tři hlavní komponenty se využívají především jako techniky zobrazení vícerozměrných dat v projekci do roviny nebo do prostoru. Výhodou je, že tato projekce zachovává vzdálenosti a úhly mezi jednotlivými objekty. V řadě případů jsou hlavní komponenty pouze jednou z fází komplexnější analýzy.

Podstata metody PCA: Zdrojová matice dat $\boldsymbol{X}$ ($n \times m$) obsahuje n objektů a m znaků. *Objekty* jsou pozorování, vzorky, experimenty, měření, pacienti, rostliny atd., zatímco *znaky* či proměnné jsou druhy signálu měření, měřená veličina, vlastnosti (sladký, kyselý, hořký, slaný, cholerický atd.), barva apod. Důležitá je zde skutečnost, že každý znak je znám pro všech n objektů. Správná skladba zdrojové matice $\boldsymbol{X}$ čili volba které znaky použít a které objekty zařadit je delikátní úkol silně závislý na charakteru každé úlohy. Velikou výhodou metody PCA je to, že lze použít jakéhokoliv počtu proměnných ve zdrojové matici $\boldsymbol{X}$ k vícerozměrné charakterizaci. Cílem každé vícerozměrné analýzy je zpracovat data tak, aby se zřetelně indikoval model a tak odkryl skrytý jev. Myšlenka sledování rozptylu je velice důležitá, protože je vlastně základním předpokladem vícerozměrné analýzy dat, že „nalezené směry maximálního rozptylu“ jsou více či méně spjaty s těmito skrytými jevy. Matematické pozadí metody je detailně popsáno v monografii [12].

Analýza shluků (CLU): Analýza shluků (Cluster analysis, CLU) patří mezi metody, které se zabývají vyšetřováním podobnosti *vícerozměrných objektů* (tj. objektů, u nichž je změřeno větší množství proměnných) a jejich klasifikací do tříd čili *shluků.* Hodí se zejména tam, kde objekty projevují přirozenou tendenci se seskupovat. Navzdory starému přísloví, že opaky se přitahují, v přírodě se ukazuje, že platí spíše pravidlo, že podobné věci se sjednocují. Nejenom ptáci podobného peří ale také ostatní živočichové, kteří sdílejí podobné či stejné vlastnosti, mají tendenci se seskupovat, shlukovat. V biologii se proto užívá shluková analýza ke klasifikování živočichů a rostlin. Tato klasifikace se nazývá numerická taxonomie. V analýze shluků je neznámá příslušnost do tříd všech objektů. Dokonce i počet tříd či shluků je neznámý.

Lze formulovat tři hlavní cíle analýzy shluků:

- *popis systematiky*, jenž je tradičním využitím shlukové analýzy pro

průzkumové cíle a taxonomii, což je empirická klasifikace objektů,
- *zjednodušení dat*, kdy analýza shluků poskytuje při hledání taxonomie zjednodušený pohled na objekty,
- *identifikaci vztahu*, kdy po nalezení shluků objektů, a tím i struktury mezi objekty, je snadnější odhalit vztahy mezi objekty.

Cíle shlukové analýzy nelze oddělit od hledání a volby vhodných znaků k charakterizování shlukovaných objektů. Nalezené shluky vystihují strukturu dat pouze s ohledem na vybrané znaky. Volba znaků musí být provedena na základě teoretických, pojmových a praktických hledisek. Vlastní shluková analýza neobsahuje techniku k rozlišení významných a nevýznamných znaků. Provede pouze odlišení shluků. Nesprávné zařazení znaků vede k zahrnutí i odlehlých objektů, které mohou mít rušivý vliv na výsledky analýzy. Měly by být využity pouze takové znaky, které dostatečně rozlišují mezi objekty.

Vliv odlehlých objektů: Při odhalování struktury objektů je shluková analýza velmi citlivá na přítomnost *nevýznamných znaků.* Je ovšem citlivá také na přítomnost *odlehlých objektů,* které se silně odlišují ode všech ostatních objektů. Odlehlé objekty mohou představovat buď skutečně odchýlené, patologické objekty, které nepředstavují analyzované populace, nebo chybný výběr objektu z populace, který způsobí nevhodné zastoupení původní populace. V obou případech odchýlené objekty zbortí strukturu dat a způsobí, že nalezené shluky nebudou odrážet skutečnou strukturu analyzované populace. Postupy průzkumové analýzy vícerozměrných dat se stávají těžkopádnými při velkém počtu objektů nebo znaků. Vybočující objekty je obvykle třeba z dat odstranit, i když si musíme být vědomi, že jejich odstraněním se mnohdy zbortí aktuální struktura. Vypouštění objektů by proto mělo být velmi uvážlivé.

Míry podobnosti: Myšlenka podobnosti objektů je v analýze shluků základní. Podobnost mezi objekty je užita jako kritérium tvorby shluků objektů. Nejdříve se stanovují znaky určující podobnost, které se dále kombinují do podobnostních měr. Tímto způsobem pak může být objekt porovnán s jiným objektem. Analýza shluků vytváří shluky podobných objektů. Meziobjektová podobnost může být měřena rozličnými způsoby, které se dají obyčejně zařadit do jedné ze tří základních skupin, a to míry korelace, míry vzdálenosti a míry asociace. Každá z nich představuje zvláštní pohled na podobnost, která je závislá na objektech a na typu dat. Korelační a vzdálenostní míry jsou míry metrických dat, zatímco asociační míry jsou určeny spíše pro nemetrická data.

Korelační míry. Základní mírou podobnosti dvou objektů či znaků x_i a x_j, vyjádřených v kardinální škále, může být *Pearsonův párový korelační koeficient* r. Objekty jsou si tím podobnější, čím je jejich párový korelační koeficient větší a bližší jedné. V případě ordinální škály je analogickou mírou podobností *Spearmanův korelační koeficient.* Obyčejně se vychází z transponované matice dat $\boldsymbol{X}^{\mathrm{T}}$, kdy sloupce představují objekty a řádky pak znaky. Korelační koeficienty mezi

dvěma sloupci matice $\boldsymbol{X}^{\mathrm{T}}$ představují korelaci mezi dvojicí objektů. Tomu odpovídá podobnost jejich profilů v profilovém diagramu. Vysoká korelace prozrazuje vysokou „podobnost“ a nízká korelace pak „nepodobnost“ profilů. Korelační míry představují podobnost odpovídajících si „vzorů“ posuzovanou přes všechny znaky. Korelační míry se však užívají zřídka, protože v praktické analýze je kladen důraz více na velikosti objektů než na tvar jejich profilových křivek.

Míry vzdálenosti. Představují nejčastěji užívané míry založené na prezentaci objektů v prostoru, jehož souřadnice tvoří jednotlivé znaky. Pokud tyto míry splňují požadavky symetrie $d(x, y) = d(y, x)$ a trojúhelníkovou nerovnost $d(x, y) \leq d(x, z) + d(y, z)$, jde o tzv. *metriky.* Při porovnání na profilovém diagramu je zřejmé, že vzdálenosti se zaměřují na velikost hodnot a vyhledání křivek v profilovém diagramu, které jsou blízko sebe, i když u jednotlivých znaků velmi odlišného tvaru. Použití korelace vede na jiné shluky než použití vzdálenostních měr. Nejčastější vzdálenostní mírou je *eukleidovská vzdálenost* zvaná také *geometrická metrika*, která představuje délku přepony pravoúhlého trojúhelníka a její výpočet je založen na Pythagorově větě. Platí, že

$$d_{\mathrm{E}}(\boldsymbol{x}_k, \boldsymbol{x}_l) = \sqrt{\sum_{j=1}^{m} (x_{kj} - x_{lj})^2}$$

představuje standardní typ vzdálenosti. Vedle eukleidovské vzdálenosti se užívá také *čtverec eukleidovské vzdálenosti,* který tvoří základ Wardovy metody shlukování. Existuje však ještě několik dalších vzdálenostních měr.

Míry asociace. Míry asociace, resp. podobnosti (binární) se používají k porovnání objektů, pokud jsou jejich znaky nemetrického charakteru (binární proměnné). Uveďme příklad, kdy respondent odpověděl na řadu otázek odpovědí *ano* nebo *ne*. Míra asociace pak vyjadřuje stupeň souhlasu každého páru respondentů. Nejjednodušší mírou asociace bude procento souhlasu, kdy oba respondenti na danou otázku odpověděli *ano* nebo *ne*. Rozšíření tohoto jednoduchého „souhlasného koeficientu“ je podstatou míry asociace k vyhodnocování více kategorií nominálních nebo ordinálních znaků. Přehled různých typů koeficientů asociace lze nalézt v pracích.

Dendrogramy hierarchického shlukování: Analýzou shluků je možné hodnotit jednak podobnost objektů, analyzovanou pomocí dendrogramu objektů, a jednak podobnost znaků analyzovanou pomocí dendrogramu znaků.

Dendrogram shluků (vývojový strom) se konstruuje pouze v případě, když je k dispozici matice původních znaků.

Dendrogram podobnosti objektů je standardní výstup hierarchických shlukovacích metod, ze kterého je patrná struktura objektů ve shlucích.

Dendrogram podobnosti znaků odhaluje nejčastěji dvojice či trojice (obecně *m*-tice) znaků, které jsou si velmi podobné a silně spolu korelují. Znaky, které jsou ve společném shluku si jsou značně podobné a jsou také vzájemně nahraditelné. To má značný význam při plánování experimentu a respektování úsporných ekonomických

kritérií. Některé vlastnosti či znaky není třeba vůbec měřit, protože jsou snadno nahraditelné jinými a nepřispívají do celku velkou vypovídací schopnost.

Míra věrohodnosti: Při sestrojování dendrogramu, k čemuž máme celou řadu technik, posuzujeme *míru věrohodnosti* nebo-li *těsnost proložení.* Prvním kritériem při volbě „nejlepšího dendrogramu“, jenž nejlépe odpovídá struktuře objektů a znaků mezi objekty, je *kofenetický korelační koeficient CC*. Je to Pearsonův korelační koeficient mezi skutečnou a predikovanou vzdáleností, založenou na dendrogramu. Čím vyšší hodnota *CC*, tím větší věrohodnost a lepší model shluků. Druhým kritériem těsnosti proložení je *kritérium delta* Δ, které měří stupeň přetvoření struktury dat spíše než stupeň podobnosti. Kritérium delta je definováno vztahem

$$\Delta_A = \left[\frac{\sum_{j<k}^{N} |d_{jk} - d_{jk}^*|^{1/A}}{\sum_{j<k}^{N} (d_{jk}^*)^{1/A}} \right]^A ,$$

kde A = 0.5 nebo 1, d_{ij} je vzdálenost v původní matici vzdáleností a d^*_{ij} značí vzdálenost získanou z dendrogramu. Je žádoucí, aby hodnoty Δ_A byly blízké nule. Řada autorů ukázala, že metoda průměrová vede obvykle k nejlepšímu dendrogramu.

Uvedené metody budou nyní ukázány na celé řadě praktických úloh.

Určení vzájemných vazeb

(a) *struktura a vazby v proměnných*
(b) *struktura a vazby v objektech*

(1) Hledání struktury v proměnných (metrická škála): faktorová analýza FA, analýza hlavních komponent PCA a shluková analýza.

(2) Hledání struktury v objektech (metrická škála): shluková analýza.

(3) Hledání struktury v objektech (metrická i nemetrická škála): vícerozměrné škálování.

(4) Hledání struktury v objektech (nemetrická škála): korespondenční analýza.

(5) Hledání lineárních vícerozměrných modelů (metrická i nemetrická škála): většina metod vícerozměrné statistické analýzy, kde závisle proměnné se uvažují jako lineárn kombinace nezávisle proměnných.

5) ***Pravidlo:*** má-li nějaký původní znak malý či dokonce žádný rozptyl, není schopen přispívat k rozlišení mezi objekty.

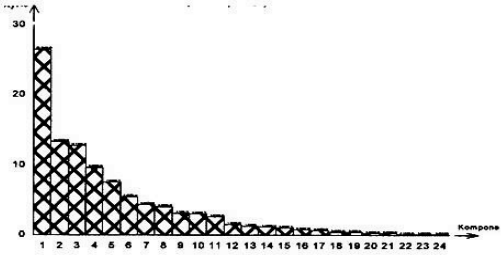
6) Využitím PCA je *snížení dimenze úlohy* čili redukce počtu znaků bez velké ztráty informace, užitím pouze prvních několika hlavních komponent.

7) *Nevyužité hlavní komponenty* obsahují malé množství informace, protože jejich rozptyl je příliš malý.

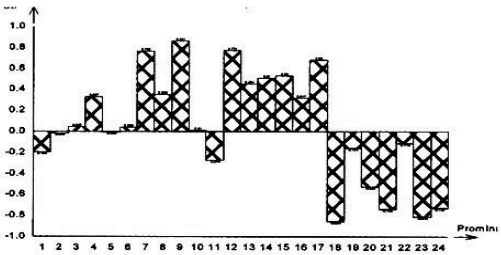
8) Hlavní komponenty jsou nekorelované.

9) První hlavní komponenta je například vhodným ukazatelem jakosti.

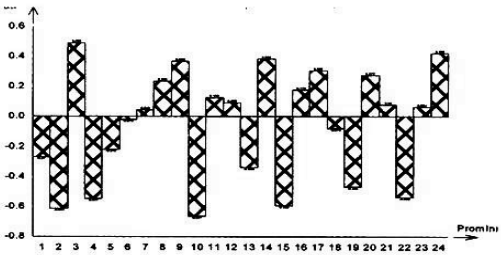
10) První dvě resp. první tři hlavní komponenty se využívají především jako techniky zobrazení vícerozměrných dat v projekci do roviny (nebo do prostoru).



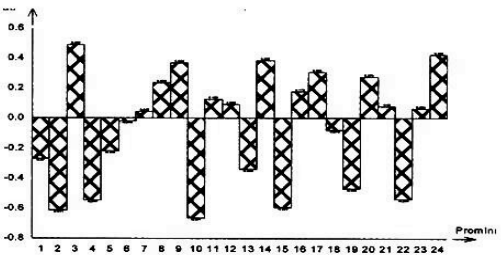
Obr. 1. Sloupcový diagram indexového grafu úpatí pro 38 objektů a 24 původních proměnných zdrojové matice Wine.



Obr. 2. Složení 1. hlavní komponenty z původních proměnných pro 38 objektů a 24 původních proměnných zdrojové matice Wine.



Obr. 3. Složení 2. hlavní komponenty z původních proměnných pro 38 objektů a 24 původních proměnných zdrojové matice Wine.



Obr. 4. Složení 3. hlavní komponenty z původních proměnných pro 38 objektů a 24 původních proměnných zdrojové matice Wine.

Výklad:

Graf komponentních vah, zátěží

(Plot Components Weights)

Zobrazí: komponentní váhy

Porovnávají se: vzdálenosti mezi proměnnými.
Krátká znamená silnou korelaci.

Nalezneme: shluk podobných proměnných, jež spolu korelují.

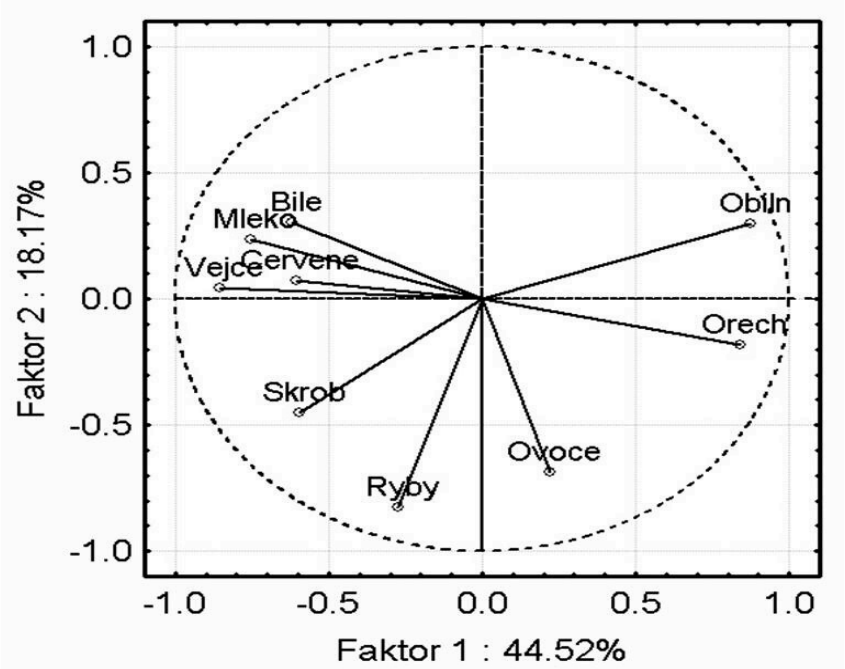
Představuje: most mezi původními proměnnými a hlavními komponentami.

Ukazuje: jakou měrou přispívají jednotlivé původní proměnné do hlavních komponent.

Pojmenovat: podaří se hlavní komponenty y_1, y_2 a vysvětlit a přidělit jim fyzikální, chemický nebo biologický význam.

Původní proměnné x_j přispívají: kladnou vahou nebo zápornou.

Sledujeme: kovarianci původních proměnných x_j v grafu komponentních vah y_1, y_2 a y_3: jsou-li proměnné x_j, $j = 1,..., m$, blízko sebe v prostorovém shluku, jde o silnou pozitivní kovarianci.



Rozptylový diagram komponentního skóre

(Scatterplot)

1. **Umístění objektů:** daleko od počátku jsou extrémy. Objekty nejblíže počátku jsou nejtypičtější.

2. **Podobnost objektů:** objekty blízko sebe si jsou podobné, daleko od sebe jsou si nepodobné.

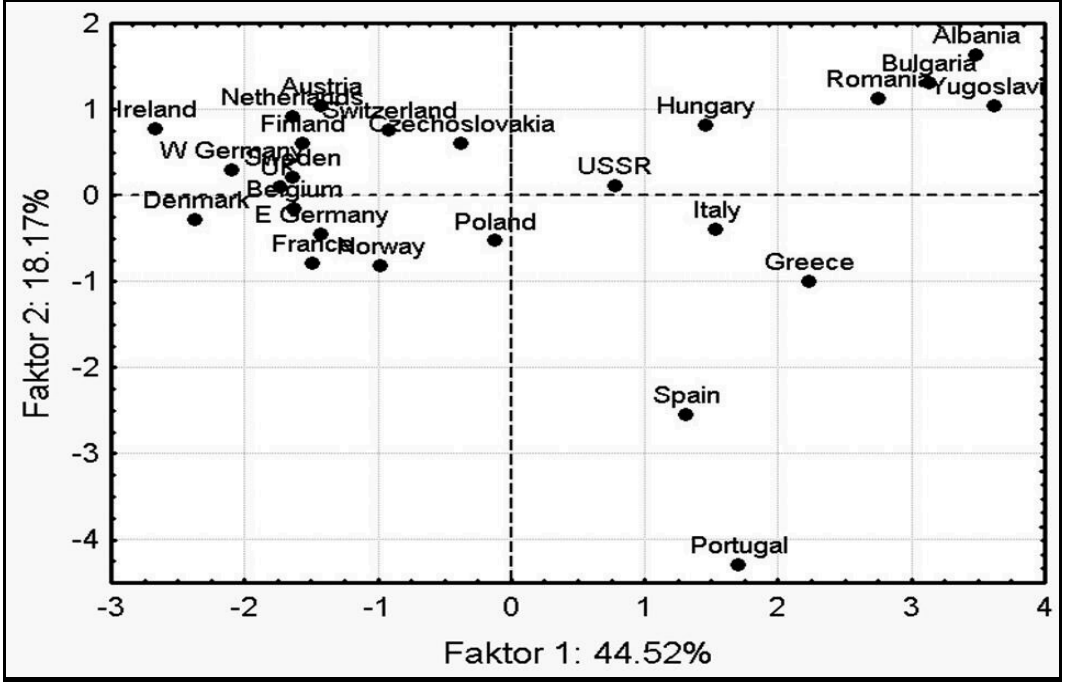
3. **Objekty v shluku:** umístěné zřetelně v jednom shluku jsou si podobné a nepodobné objektům v ostatních shlucích. Jsou-li shluky blízko sebe, znamená to značnou podobnost objektů.

4. **Osamělé objekty:** izolované objekty mohou být odlehlé.

5. **Odlehlé objekty:** ideálně bývají objekty rozptýlené po celé ploše diagramu. V opačném případě je špatný model.

6. **Pojmenování objektů:** výstižná jména objektů slouží k hledání hlubších souvislostí mezi objekty a vystihneme tak jejich fyzikální či biologický vztah.

7. **Vysvětlení místa objektu:** umístění objektu na ploše v diagramu může být porovnáváno s komponentními vahami původních proměnných ve dvojném grafu.



Výklad:

Indexový graf úpatí vlastních čísel

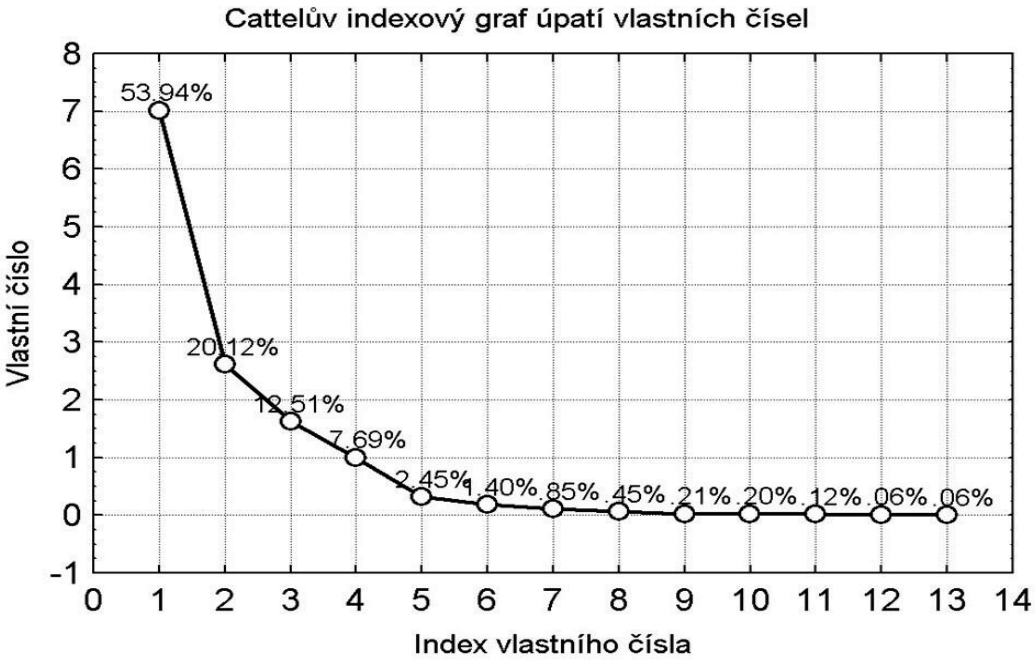
(Scree Plot)

Je to sloupcový diagram vlastních čísel proti indexu *A*.

Zobrazuje: relativní velikost jednotlivých vlastních čísel.

Využití: k určení počtu *A* "užitečných" hlavních komponent. Graf úpatí se jeví nejobjektivnějším kritériem.

Kritérium "1": hrubším kritériem PC, jejichž vlastní číslo je větší než jedna. Graf úpatí se však jeví objektivnějším.



Diagnostika metody PCA

Maticový graf rozptylových diagramů znaků slouží k získání počáteční informace o datech, zda data potřebují škálování. V PCA postupně provádíme:

1. ***Vyšetření indexového grafu úpatí vlastních čísel*** – z hrany úpatí v tomto diagramu se určí vhodný počet hlavních komponent.

2. ***Výpočet vlastních vektorů*** – vedle číselných hodnot se užívá i názorný čárový diagram hodnot vlastních vektorů, který přehledně informuje o relativním zastoupení původních znaků x_j , $j = 1, ..., m$, v hlavních komponentách.

3. ***Výpočet komponentních vah*** – matice párových korelačních koeficientů obsahující korelace původních znaků s hlavními komponentami. Uživatel nyní vybere pouze prvních k hlavních komponent a vytvoří tak model PCA.

4. ***Vyšetření grafu komponentních vah***.

5. ***Vyšetření rozptylového diagramu komponentního skóre.***

6. ***Vyšetření dvojného grafu***.

7. ***Vyšetření reziduí*** – rezidua objektů a rezidua proměnných by měla prokazovat dostatečnou těsnost proložení.

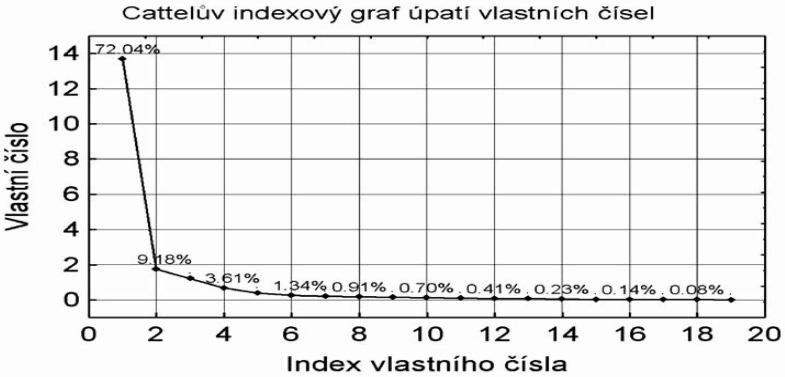
8. ***Určení významných původních znaků*** – je výhodné vyhledávat významné znaky, protože klasická metoda PCA umožňuje sice redukci počtu hlavních komponent, ale každá komponenta zůstává stále kombinací všech původních znaků.

Úloha 1. Klasifikace polétavých mšic (Kompendium B404)

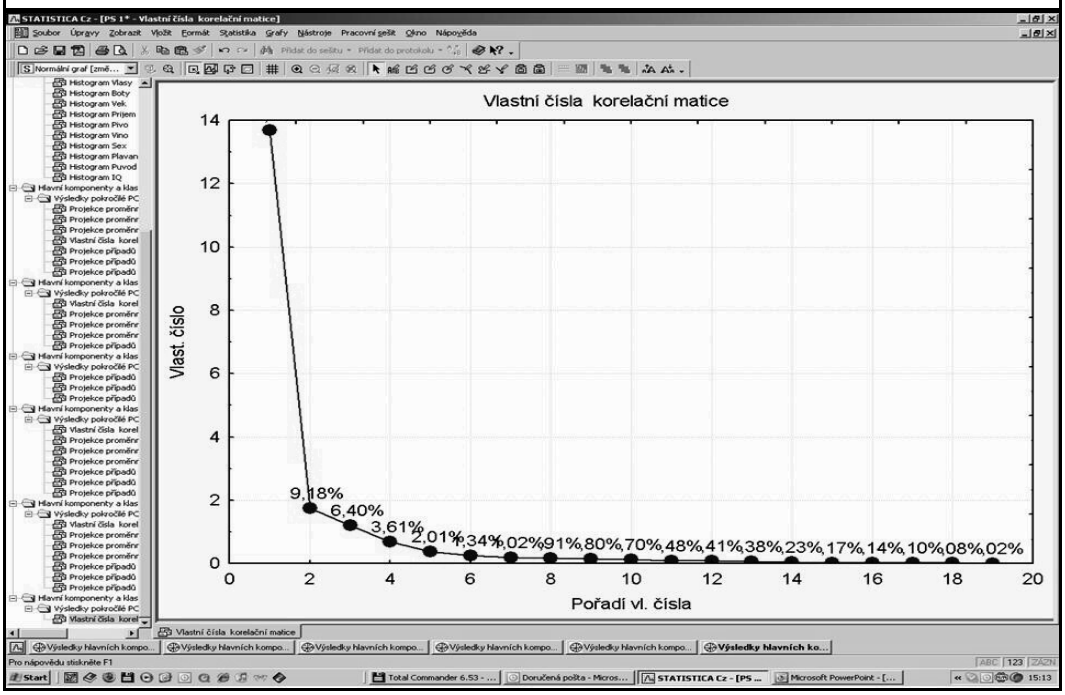
Jeffers (1967)25 studoval 40 jedinců polétavých mšic *(Alate adelges)*: 19 ukazatelů k rozlišení druhů, 14 znaků délky a šířky, 4 znaky se týkají počtu a 1 binární vyjadřuje přítomnost či absenci: x1 délka těla, x2 šířka těla, x3 délka předního křídla, x4 délka zadního křídla, x5 počet průduchů, x6 délka tykadla I, x7 délka tykadla II, x8 délka tykadla III, x9 délka tykadla IV, x10 délka tykadla V, x11 počet tykadlových ostnů, x12 délka posledního článku nohy, x13 délka holeně, tibia, x14 délka stehna, x15 délka sosáku, x16 délka kladélka, x17 počet kladélkových trnů, x18 řitní otvor, x19 počet háčků zadních křídel

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18	x19
21.2	11	7.5	4.8	5	2	2	2.8	2.8	3.3	3	4.4	4.5	3.6	7	4	8	0	3
20.2	10	7.5	5	5	2.3	2.1	3	3	3.2	5	4.2	4.5	3.5	7.6	4.2	8	0	3
20.2	10	7	4.6	5	1.9	2.1	3	2.5	3.3	1	4.2	4.4	3.3	7	4	6	0	3
22.5	8.8	7.4	4.7	5	2.4	2.1	3	2.7	3.5	5	4.2	4.4	3.6	6.8	4.1	6	0	3
20.6	11	8	4.8	5	2.4	2	2.9	2.7	3	4	4.2	4.7	3.5	6.7	4	6	0	3
19.1	9.2	7	4.5	5	1.8	1.9	2.8	3	3.2	5	4.1	4.3	3.3	5.7	3.8	8	0	3.5
20.8	11.4	7.7	4.9	5	2.5	2.1	3.1	3.1	3.2	4	4.2	4.7	3.6	6.6	4	8	0	3
15.5	8.2	6.3	4.9	5	2	2	2.9	2.4	3	3	3.7	3.8	2.9	6.7	3.5	6	0	3.5
16.7	8.8	6.4	4.5	5	2.1	1.9	2.8	2.7	3.1	3	3.7	3.8	2.8	6.1	3.7	8	0	3
19.7	9.9	8.2	4.7	5	2.2	2	3	3	3.1	0	4.1	4.3	3.3	6	3.8	8	0	3
10.6	5.2	3.9	2.3	4	1.2	1	2	2	2.2	6	2.5	2.5	2	4.5	2.7	4	1	2
9.2	4.5	3.7	2.2	4	1.3	1.2	2	1.6	2.1	5	2.4	2.3	1.8	4.1	2.4	4	1	2
9.6	4.5	3.6	2.3	4	1.3	1	1.9	1.7	2.2	4	2.4	2.3	1.7	4	2.3	4	1	2
8.5	4	3.8	2.2	4	1.3	1.1	1.9	2	2.1	5	2.4	2.4	1.9	4.4	2.3	4	1	2
11	4.7	4.2	2.3	4	1.2	1	1.9	2	2.2	4	2.5	2.5	2	4.5	2.6	4	1	2
18.1	8.2	5.9	3.5	5	1.9	1.9	1.9	2.7	2.8	4	3.5	3.8	2.9	6	4.5	9	1	2
17.6	8.3	6	3.8	5	2	1.9	2	2.2	2.9	3	3.5	3.6	2.8	5.7	4.3	10	1	2
19.2	6.6	6.2	3.4	5	2	1.8	2.2	2.3	2.8	4	3.5	3.4	2.5	5.3	3.8	10	1	2
15.4	7.6	7.1	3.4	5	2	1.9	2.5	2.5	2.9	4	3.3	3.6	2.7	6	4.2	8	1	3
15.1	7.3	6.2	3.8	5	2	1.8	2.1	2.4	2.5	4	3.7	3.7	2.8	6.4	4.3	10	1	2.5
16.1	7.9	5.8	3.7	5	2.1	1.9	2.3	2.6	2.9	5	3.6	3.6	2.7	6	4.5	0	1	2
19.1	8.8	6.4	3.9	5	2.2	2	2.3	2.4	2.9	4	3.8	4	3	6.5	4.5	0	1	2.5
15.3	6.4	5.3	3.3	5	1.7	1.6	2	2.2	2.5	5	3.4	3.4	2.6	5.4	4	0	1	2
14.8	8.1	6.2	3.7	5	2.2	2	2.2	2.4	3.2	5	3.5	3.7	2.7	6	4.1	0	1	2
16.2	7.7	6.9	3.7	5	2	1.8	2.3	2.4	2.8	4	3.8	3.7	2.7	5.7	4.2	0	1	2.5
13.4	6.9	5.7	3.4	5	2	1.8	2.8	2	2.6	4	3.6	3.6	2.6	5.5	3.9	0	1	2
12.9	5.8	4.8	2.6	5	1.6	1.5	1.9	2.1	2.6	5	2.8	3	2.2	5.1	3.6	9	1	3
12	6.5	5.3	3.2	5	1.9	1.9	2.3	2.5	3	5	3.3	3.5	2.6	5.4	4.3	8	1	2
14.1	7	5.5	3.6	5	2.2	2	2.3	2.5	3.1	5	3.6	3.7	2.8	5.8	4.1	0	1	2
16.7	7.2	5.7	3.5	5	1.9	1.9	2.5	2.3	2.8	5	3.4	3.6	2.7	6	4	0	1	2.5
14.1	5.4	5	3	5	1.7	1.6	1.8	2.5	2.4	5	2.7	2.9	2.2	5.3	3.6	8	1	2
10	6	4.2	2.5	5	1.6	1.4	1.4	2	2.7	6	2.8	2.5	1.8	4.8	3.4	8	1	2
11.4	4.5	4.4	2.7	5	1.8	1.5	1.9	1.7	2.5	5	2.7	2.5	1.9	4.7	3.7	8	1	2
12.5	5.5	4.7	2.3	5	1.8	1.4	1.8	2.2	2.4	4	2.8	2.6	2	5.1	3.7	8	0	2
13	5.3	4.7	2.3	5	1.6	1.4	1.8	1.8	2.5	4	2.7	2.7	2.1	5	3.6	8	1	2
12.4	5.2	4.4	2.6	5	1.6	1.4	1.8	2.2	2.2	5	2.7	2.5	2	5	3.2	6	1	2
12	5.4	4.9	3	5	1.7	1.5	1.7	1.9	2.4	5	2.7	2.7	2	4.2	3.7	6	1	2
10.7	5.6	4.5	2.8	5	1.8	1.4	1.8	2.2	2.4	4	2.7	2.6	2	5	3.5	8	1	2
11.1	5.5	4.3	2.6	5	1.7	1.5	1.8	1.9	2.4	5	2.6	2.5	1.9	4.6	3.4	8	1	2
12.8	5.7	4.8	2.8	5	1.6	1.4	1.7	1.9	2.3	5	2.3	2.5	1.9	5	3.1	8	1	2

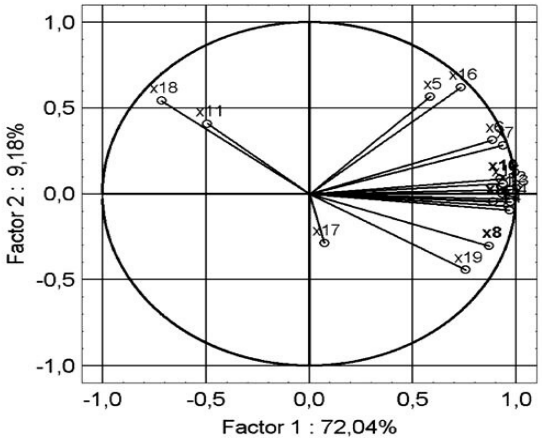
1. Cattelův indexový graf úpatí vlastních čísel: z 19 znaků lze snížit rozměrnost na první dvě hlavní komponenty, které popisují přes 81% původní proměnlivosti v datech.



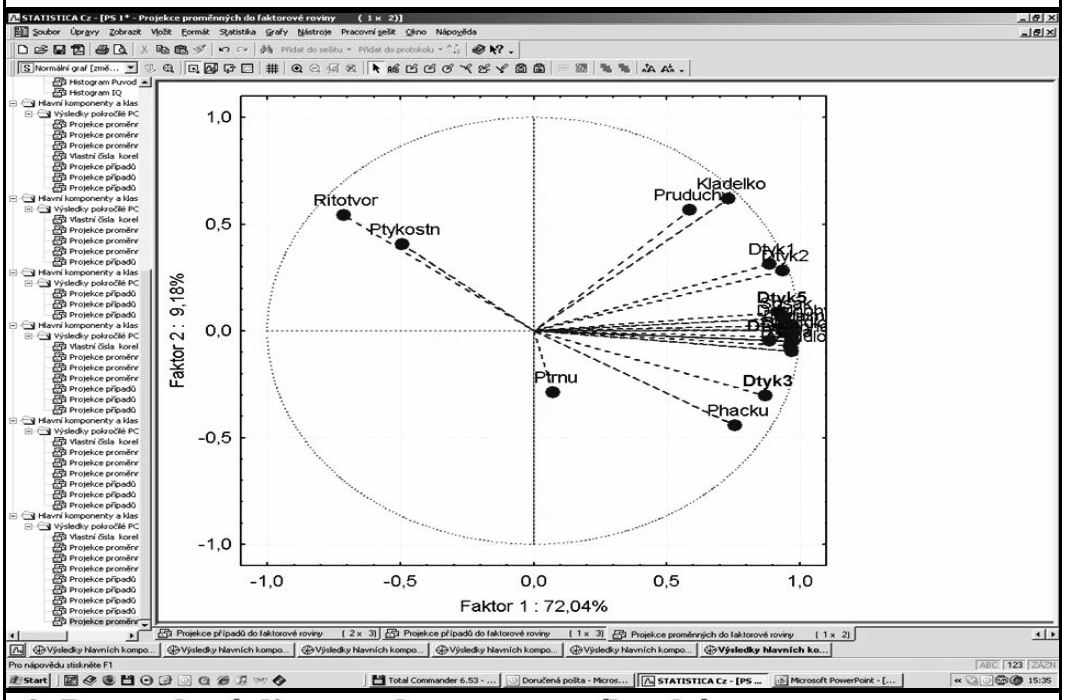
Obr. 4.23 Cattelův indexový graf úpatí vlastních čísel Scree plot dat *Mšice* (STATISTICA).



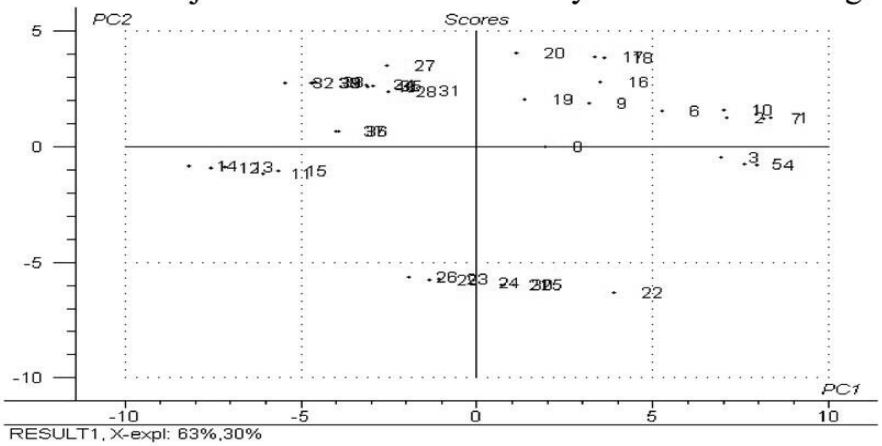
2. Graf komponentních vah: roztřídí 19 znaků: vedle shluku společných znaků jsou x_1 a x_{17} odlehlé od ostatních. Od shluku jsou odděleny znaky x_2 a x_3, a dále x_{11} a x_{13}. Znaky x_2 a x_3 spolu pozitivně korelují, dále x_{11} a x_{18} spolu pozitivně korelují ale negativně korelují se x_1, x_2 a x_3. Znak x_1 pozitivně koreluje s x_2, a x_1 koreluje s x_3.



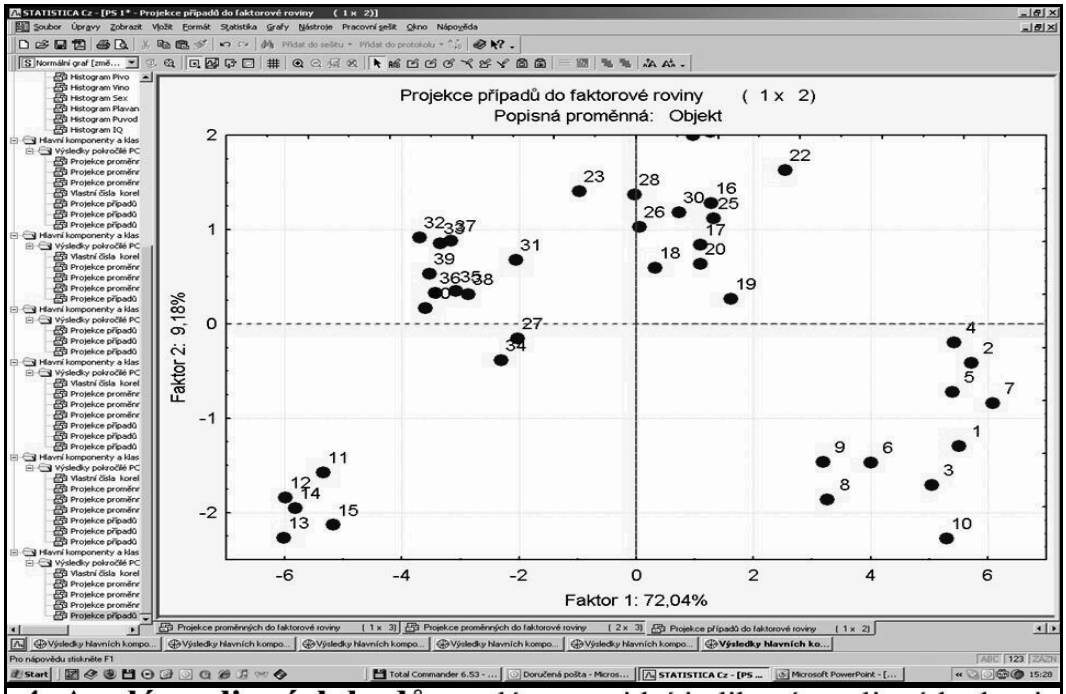
Obr. 4.24 Graf komponentních vah 1 a 2 zdrojové matice dat *Mšice* (STATISTICA).



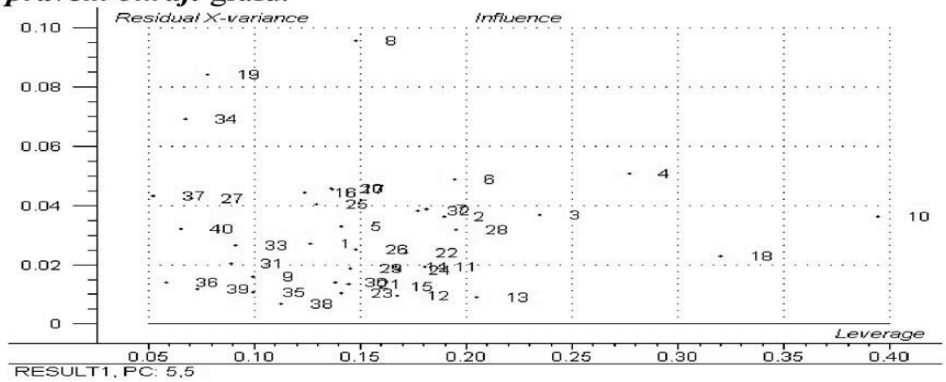
3. Rozptylový diagram komponentního skóre: mšice jsou roztříděny do 4 shluků. Závěr je v souhlase se taxonomických tříděním z biologie.



Obr. 4.25 Rozptylový diagram komponentního skóre dat *Mšice* (UNSCRAMBLER).



4. Analýza vlivných bodů: analýzou reziduí indikovány vlivné body, tj. ***odlehlé objekty*** nesouhlasící s navrženým modelem PCA při ***horním okraji*** grafu, a ***extrémní objekty,*** které souhlasí s navrženým modelem PCA a jsou při ***pravém okraji*** grafu.



Obr. 4.26 Graf vlivných bodů statistické analýzy reziduí dat *Mšice* (UNSCRAMBLER).

○ *Závěr:* PCA je užitečná při taxonomickém třídění mšic: nalezeny 4 shluky mšic.

PŘÍKLAD 9.4 *Vytvoření dendrogramu neuroleptik*

Neuroleptika redukují nežádoucí účinky přebytečného dopaminu a liší se ve svých účincích: potlačují nervozitu, záchvaty, třes, ospalost, parkinsonismus, vynechávání menstruace, vyrážky, zvýšené slinění atd. Cílem je provést klasifikaci neuroleptik do shluků podobných účinků.

○ ***Data*:** Data ***Neuroleptika*** (převrácená hodnota mediánové účinné dávky *1/ED50* [kg/mg]):
Lek název neuroleptika,
Nervoz potlačení nervozity,
Stereo potlačení stereotypního chování,
Tres potlačení záchvatu a třesu a
Usmr dávka smrtícího účinku.

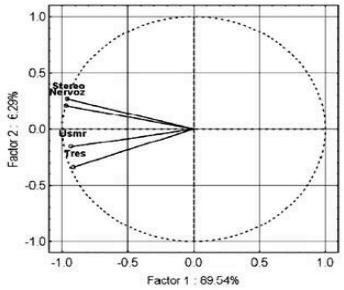
Lek	*Nervoz*	*Stereo*	*Tres*	*Usmr*
1 Chlorphromazine	3.846	3.333	1.111	1.923
2 Promazine	0.323	0.213	0.108	1.429
3 Trifluperazine	27.027	17.857	0.562	0.14
4 Fluphenazine	17.857	15.385	1.695	1.075
5 Perphenazine	27.027	27.027	1.961	2.083
6 Thioridazine	0.244	0.185	0.093	1.333
7 Pifluthixol	142.857	142.857	20.408	163.934
8 Thiothixene	4.348	4.348	0.047	0.345
9 Chorprothixene	5.882	2.941	4.545	4.167
10 Spiperone	62.5	47.619	11.765	0.847
11 Haloperidol	52.632	62.5	1.282	0.568
12 Azaperone	2.941	1.282	2.222	3.03
13 Pipamperone	0.327	0.187	1.724	0.397
14 Pimozide	20.408	20.408	0.107	0.025
15 Metitepine	15.385	10.204	10.204	27.027
16 Clozapine	0.161	0.093	0.327	0.323
17 Perlapine	0.323	0.323	0.37	0.067
18 Sulpiride	0.047	0.047	0.003	0.001
19 Butaclamol	10.204	9.091	1.471	0.025
20 Molindone	7.692	7.692	0.14	38138

○ ***Řešení:*** Po vyhledání optimální tvorby dendrogramu sestrojíme dendrogram podobnosti znaků a dendrogram podobnosti objektů.

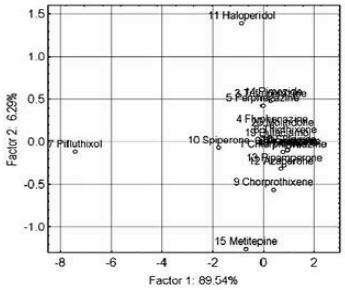
Nejvyšší hodnota kofenetického korelačního koeficientu ***CC*** a nejnižší hodnota obou kritérií delta, ***Delta(0.5)*** a ***Delta(1.0)***, vybrala **metodu skupinového průměru** (software NCSS2004).

1. Nejbližšího souseda, *Kofenetická korelace CC:* 0.988598, *Delta(0.5):* 0.474238, *Delta(1.0):* 0.391993.
2. Nejvzdálenějšího souseda: *Kofenetická korelace CC:* 0.982795, *Delta(0.5):* 0.178589, *Delta(1.0):* 0.183477;
3. Párový průměr, *Kofenetická korelace CC:* 0.988876, *Delta(0.5):* 0.177810, *Delta(1.0):* 0.188781;
4. **Skupinový průměr**, *Kofenetická korelace CC:* 0.987356, *Delta(0.5):* 0.137455, *Delta(1.0):* 0.125290;
5. Těžiště, *Kofenetická korelace CC:* 0.984750, *Delta(0.5):* 0.175238, *Delta(1.0):* 0.166599;
6. Median, *Kofenetická korelace CC:* 0.984215, *Delta(0.5):* 0.452308, *Delta(1.0):* 0.428346;
7. Wardova metoda, *Kofenetická korelace CC:* 0.979285, *Delta(0.5):* 0.549394, *Delta(1.0):* 0.492716.

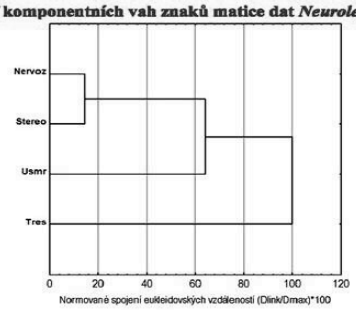
Metoda skupinového průměru v dendrogramu podobnosti objektů:
první shluk obsahuje 12 objektů 1, 8, 12, 9, 2, 6, 16, 17, 18, 13, 19, 20,
druhý shluk 5 objektů 3, 4, 14, 5, 15,
třetí shluk 2 objekty 10 a 11,
čtvrtý shluk obsahuje jeden objekt, a to 7.



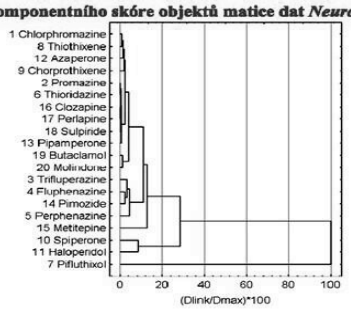
Graf komponentních vah znaků matice dat *Neuroleptika*.



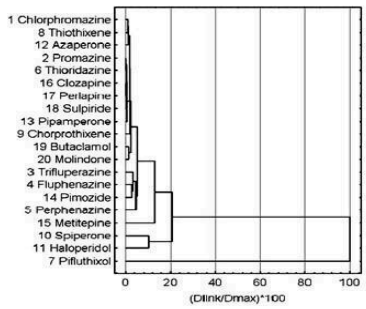
Graf komponentního skóre objektů matice dat *Neuroleptika*



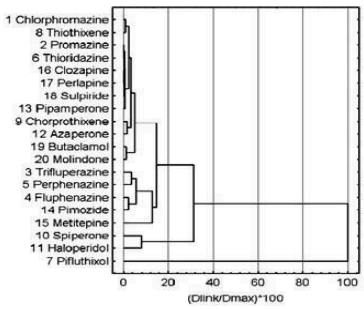
Dendrogram znaků metodou skupinového průměru



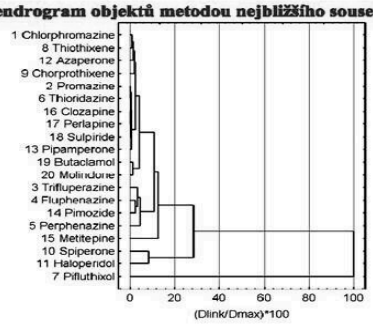
Dendrogram objektů metodou skupinového průměru



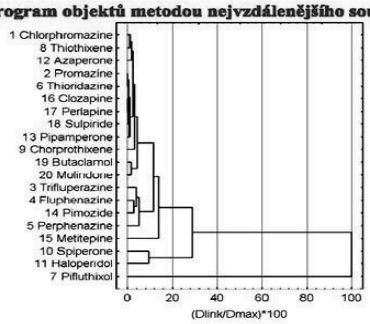
Dendrogram objektů metodou nejbližšího souseda



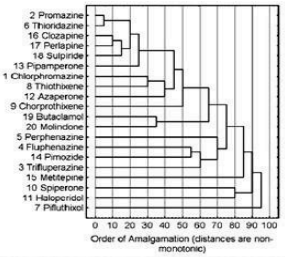
Dendrogram objektů metodou nejvzdálenějšího souseda



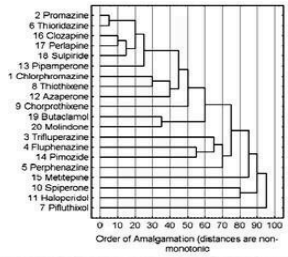
Dendrogram objektů metodou párového průměru



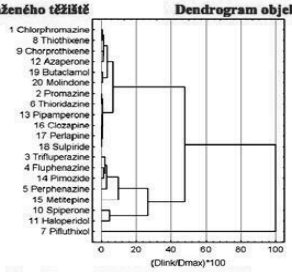
Dendrogram objektů metodou skupinového průměru



Dendrogram objektů metodou neváženého těžiště

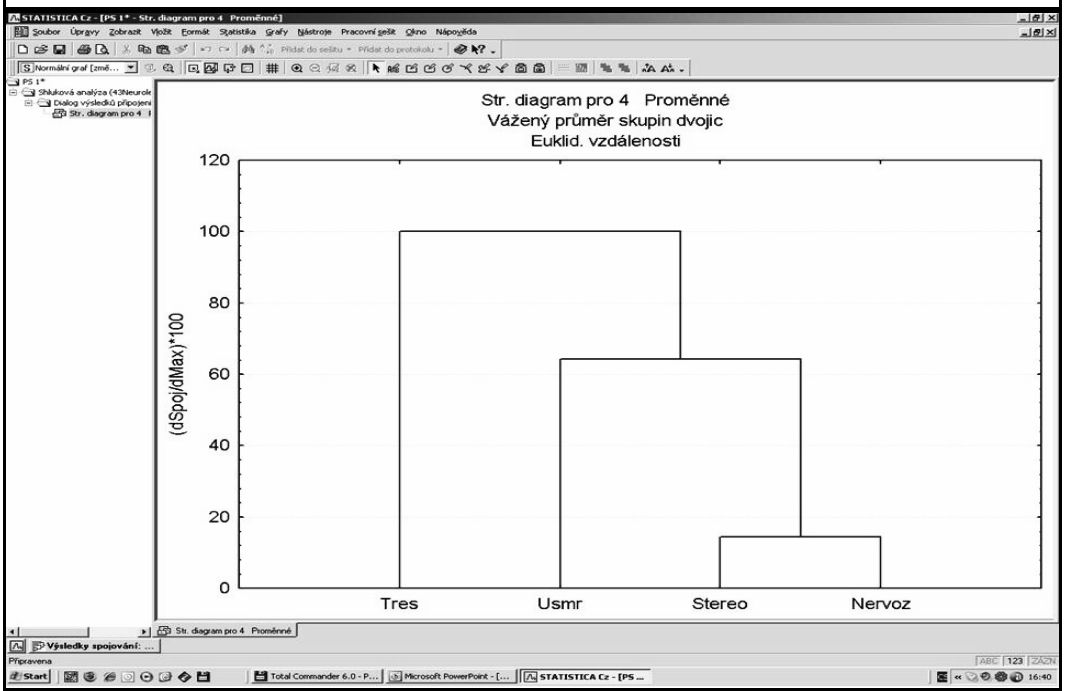


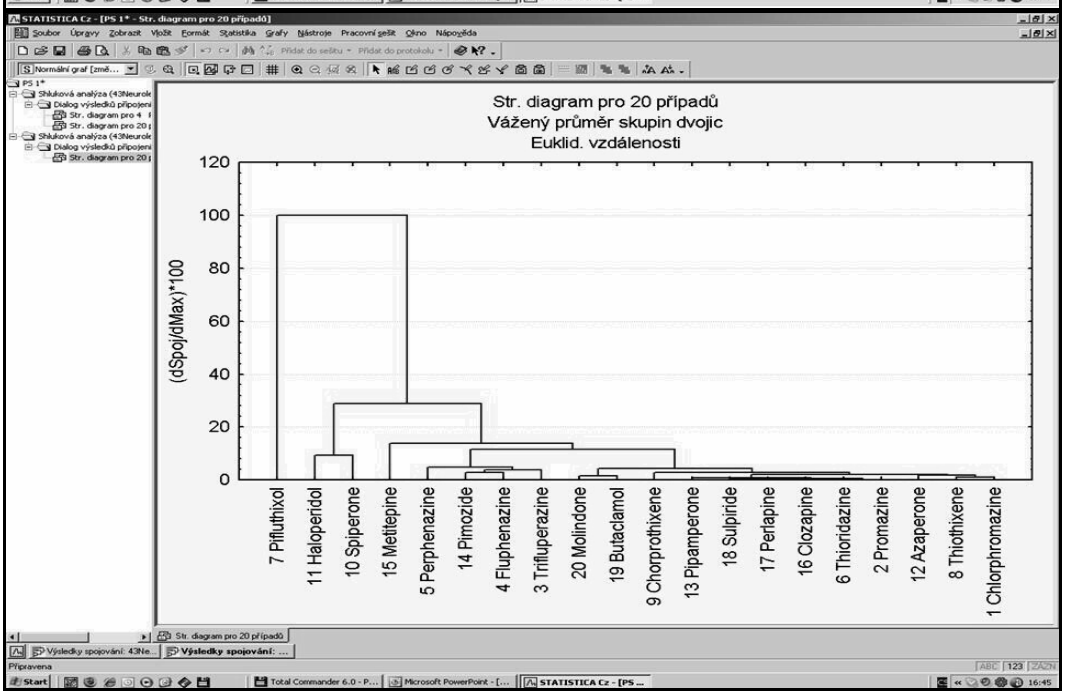
Dendrogram objektů metodou váženého těžiště (mediánu).



Dendrogram objektů metodou Wardovou

Závěr: **Nejvhodnější tvorba dendrogramu je metodami párového průměru a skupinového průměru.**





Úloha 3. Sledování spotřeby proteinů v Evropě (Kompendium B418)

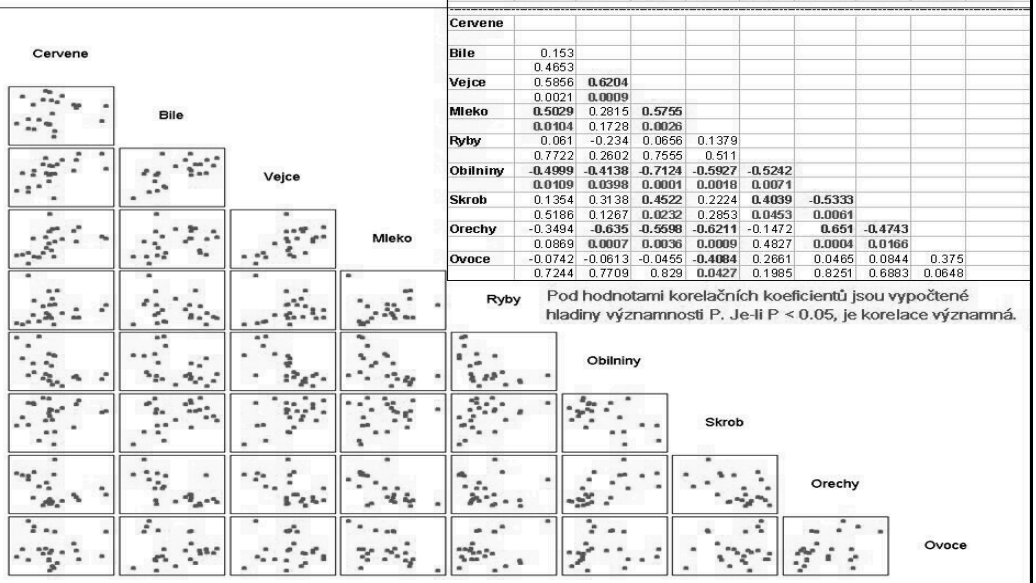
Sledovaná spotřeba proteinů v 25 zemích formou spotřeby 9 druhů potravin je předmětem vyšetření.

Data: *i* značí index, ***Cervene*** udává červené maso, **Bile** maso, **Vejce**, **Mleko**, **Ryby**, **Obilniny**, **Skrob**, **Orechy**, **Ovoce** a zelenina

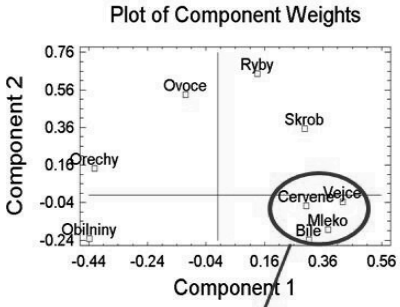
	Objekty	Proměnné								
i	**Stát**	**Cervene**	**Bile**	**Vejce**	**Mleko**	**Ryby**	**Obilniny**	**Skrob**	**Orechy**	**Ovoce**
1	Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
2	Austria	8.9	14	4.3	19.9	2.1	28	3.6	1.3	4.3
3	Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4
4	Bulgaria	7.8	6	1.6	8.3	1.2	56.7	1.1	3.7	4.2
5	Czechoslov.	9.7	11.4	2.8	12.5	2	34.3	5	1.1	4
6	Denmark	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4
7	E Germany	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
8	Finland	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1	1.4
9	France	18	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
10	Greece	10.2	3	2.8	17.6	5.9	41.7	2.2	7.8	6.5
11	Hungary	5.3	12.4	2.9	9.7	0.3	40.1	4	5.4	4.2
12	Ireland	13.9	10	4.7	25.8	2.2	24	6.2	1.6	2.9
13	Italy	9	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
14	Netherlands	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
15	Norway	9.4	4.7	2.7	23.3	9.7	23	4.6	1.6	2.7
16	Poland	6.9	10.2	2.7	19.3	3	36.1	5.9	2	6.6
17	Portugal	6.2	3.7	1.1	4.9	14.2	27	5.9	4.7	7.9
18	Romania	6.2	6.3	1.5	11.1	1	49.6	3.1	5.3	2.8
19	Spain	7.1	3.4	3.1	8.6	7	29.2	5.7	5.9	7.2
20	Sweden	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2
21	Switzerland	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
22	UK	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
23	USSR	9.3	4.6	2.1	16.6	3	43.6	6.4	3.4	2.9
24	W Germany	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
25	Yugoslavia	4.4	5	1.2	9.5	0.6	55.9	3	5.7	3.2

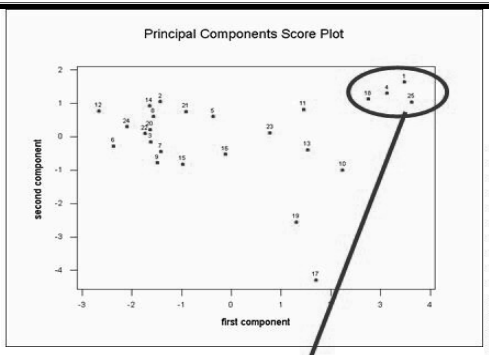
Test významnosti korelace v korelační matici

	Cervene	Bile	Vejce	Mleko	Ryby	Obilniny	Skrob	Orechy	Ovoce
Cervene									
Bile	0.153								
	0.4653								
Vejce	0.5856	**0.6204**							
	0.0021	**0.0009**							
Mleko	**0.5029**	0.2815	**0.5755**						
	0.0104	0.1728	**0.0026**						
Ryby	0.061	-0.234	0.0656	0.1379					
	0.7722	0.2602	0.7555	0.511					
Obilniny	**-0.4999**	**-0.4138**	**-0.7124**	**-0.5927**	**-0.5242**				
	0.0109	**0.0398**	**0.0001**	**0.0018**	**0.0071**				
Skrob	0.1354	0.3138	**0.4522**	0.2224	**0.4039**	**-0.5333**			
	0.5186	0.1267	**0.0232**	0.2853	**0.0453**	**0.0061**			
Orechy	-0.3494	**-0.635**	**-0.5598**	**-0.6211**	-0.1472	**0.651**	**-0.4743**		
	0.0869	**0.0007**	**0.0036**	**0.0009**	0.4827	**0.0004**	**0.0166**		
Ovoce	-0.0742	-0.0613	-0.0455	**-0.4084**	0.2661	0.0465	0.0844	0.375	
	0.7244	0.7709	0.829	**0.0427**	0.1985	0.8251	0.6883	0.0648	

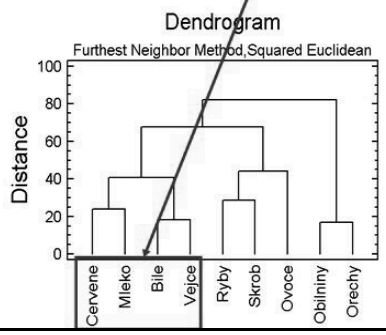


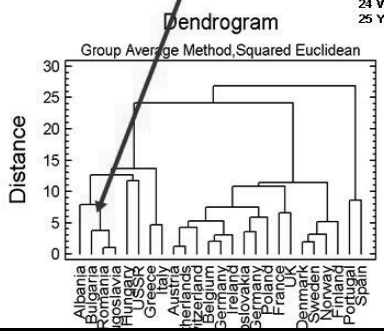
Pod hodnotami korelačních koeficientů jsou vypočtené hladiny významnosti P. Je-li P < 0.05, je korelace významná.





1 Albania
2 Austria
3 Belgium
4 Bulgaria
5 Czechoslovakia
6 Denmark
7 E Germany
8 Finland
9 France
10 Greece
11 Hungary
12 Ireland
13 Italy
14 Netherlands
15 Norway
16 Poland
17 Portugal
18 Romania
19 Spain
20 Sweden
21 Switzerland
22 UK
23 USSR
24 W Germany
25 Yugoslavia



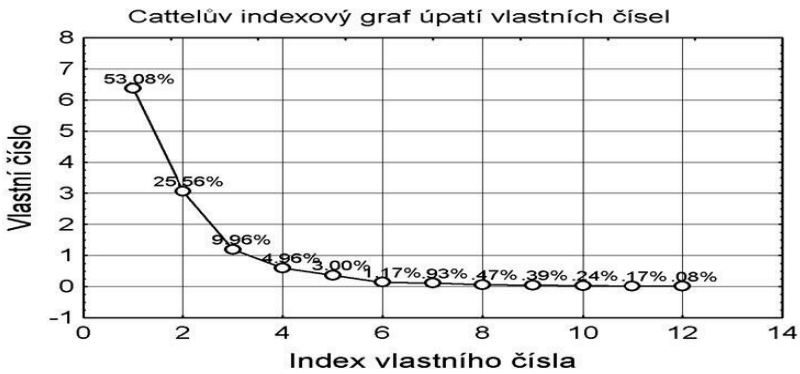


PŘÍKLAD 4.2 ***Posouzení hrachu diagramem komponentního skóre***

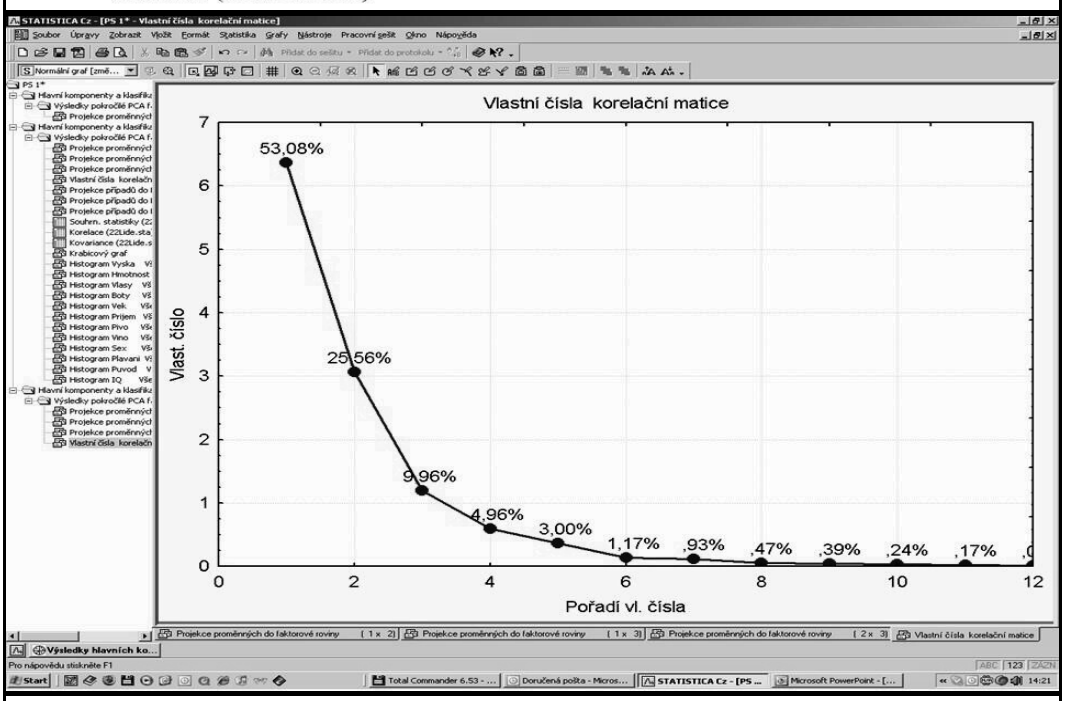
Je třeba roztřídit druhy vyšetřovaného hrachu dle smyslového posouzení hrachu člověkem, které znaky subjektivního posouzení se nejlépe hodí k popisu. Které znaky se nejlépe podílejí na popisu proměnlivosti hrachu?

○ *Řešení:*

1. Počet potřebných hlavních komponent: První hlavní komponenta popisuje 53% celkového rozptylu, druhá hlavní komponenta 25.6% a třetí hlavní komponenta 9.9%.



Obr. 4.7a Cattelův indexový graf úpatí vlastních čísel Scree Plot zdrojové matice dat *Hrách* (STATISTICA).



2. Graf komponentních vah: vysvětlení grafu

1) Vůně hrachu (znak *Aro*), sladkost (*Slad*) a medovost (*Med*) pozitivně korelují,

2) Tvrdost hrachu (*Tvrd)*, klasovost (*Klas*) a bezchuťovost (*Bez*) jsou rovněž pozitivně korelovány ale jsou negativně korelovány se znaky vůně hrachu (*Aro*), sladkost (*Slad*) a medovost (*Med*), protože oba shluky znaků leží na opačných stranách vůči počátku.

3) Druhá hlavní komponenta *PC*2 ukazuje, že barva 1 (*Bar1*), bělost (*Bel*) a ztráta (*Ztr*) jsou v horní části diagramu a obě jsou negativně korelovány s barvou 2 (*Bar2*) a barvou 3 (*Bar3*), které jsou umístěny v dolní části diagramu.

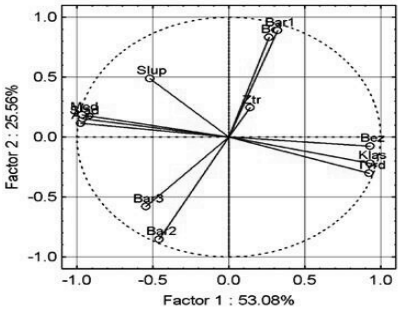
4) Vzorky hrachu nahoře diagramu jsou bělejší a vzorky v dolní části budou barevnější.

5) Slupka zrn *Slup* hrachu nekoreluje ani s bělostí (*Bel*) ani s chuťovými vlastnostmi hrachu vůně (*Aro*), sladkost (*Slad*) a medovost (*Med*).

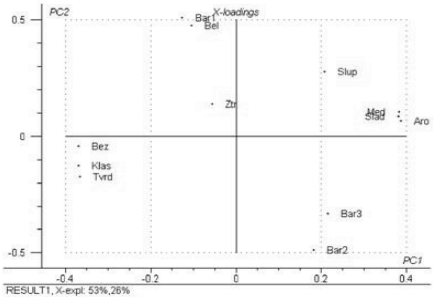
3) Druhá hlavní komponenta *PC*2 ukazuje, že barva 1 (*Bar1*), bělost (*Bel*) a ztráta (*Ztr*) jsou v horní části diagramu a obě jsou negativně korelovány s barvou 2 (*Bar2*) a barvou 3 (*Bar3*), které jsou umístěny v dolní části diagramu.

4) Vzorky hrachu nahoře diagramu jsou bělejší a vzorky v dolní části budou barevnější.

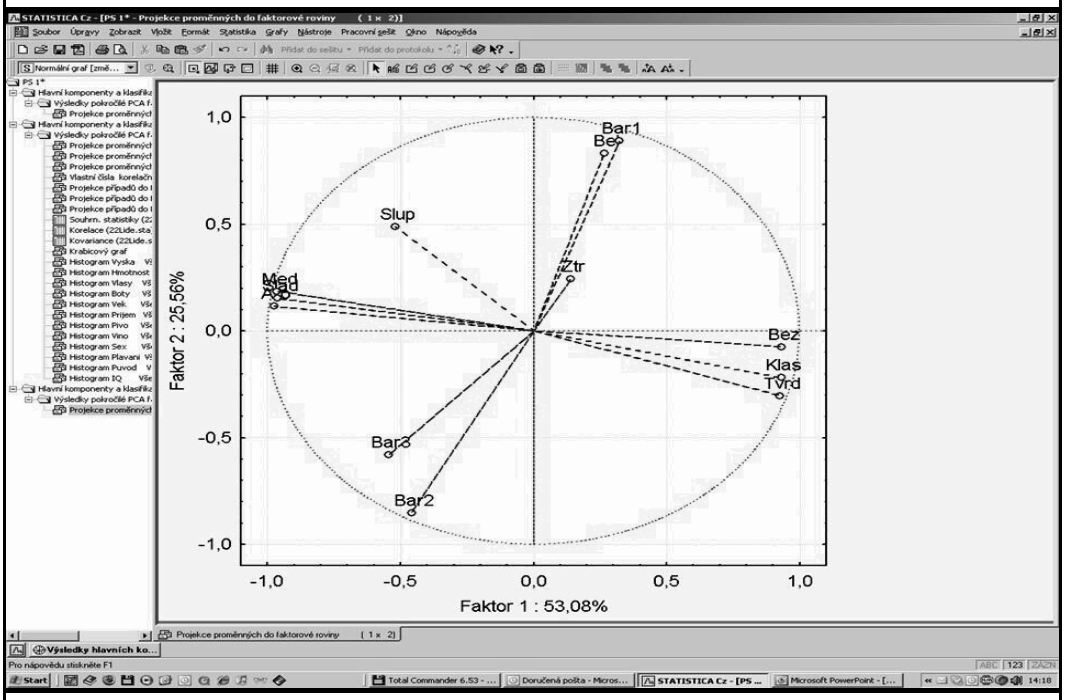
5) Slupka zrn *Slup* hrachu nekoreluje ani s bělostí (*Bel*) ani s chuťovými vlastnostmi hrachu vůně (*Aro*), sladkost (*Slad*) a medovost (*Med*).



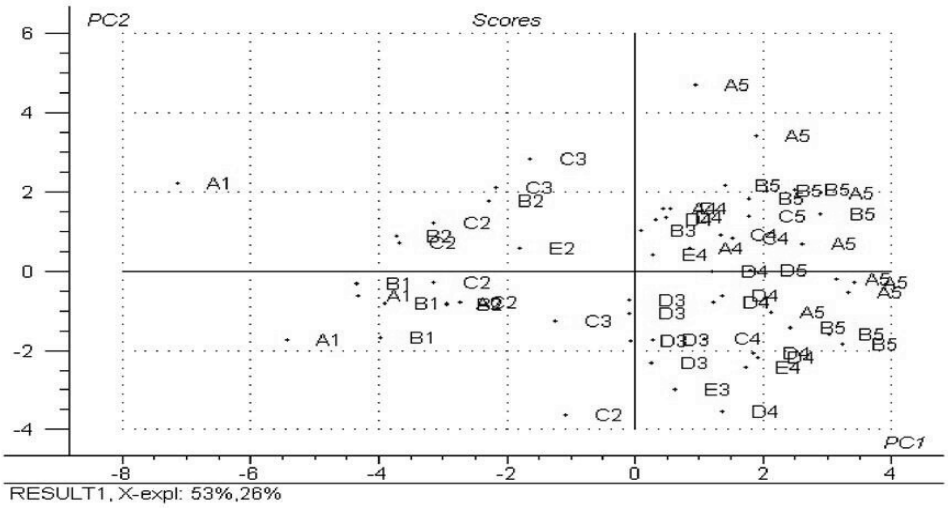
Obr. 4.8a Graf komponentních vah 1 a 2 matice dat *Hrách*.



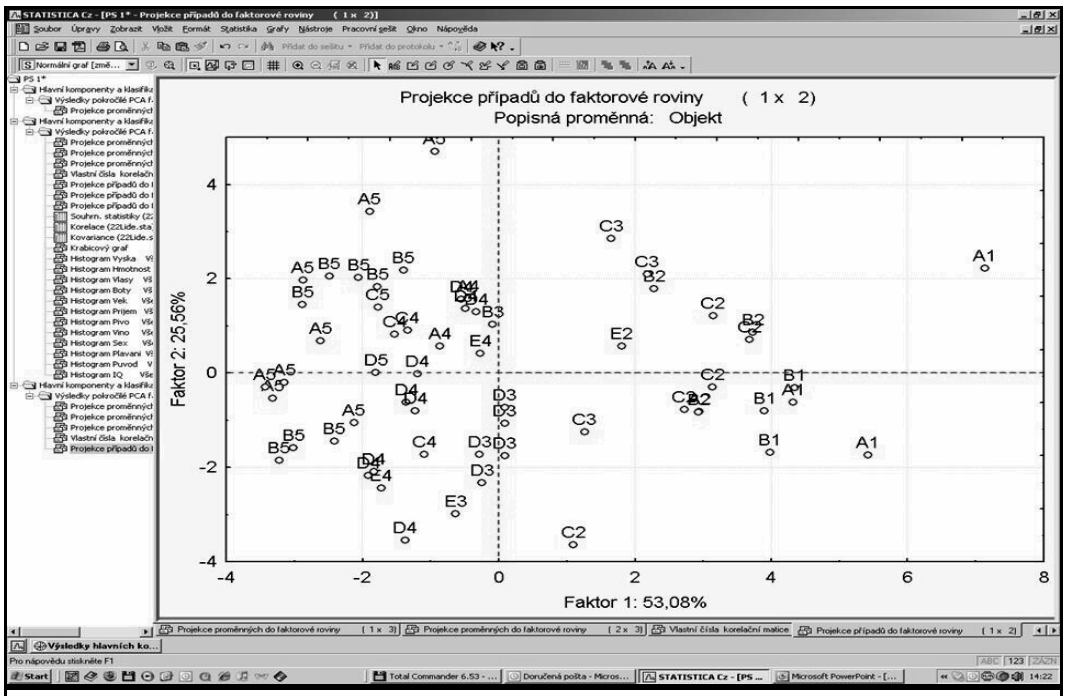
Obr. 4.8b Graf komponentních vah 1 a 2 matice dat *Hrách*.



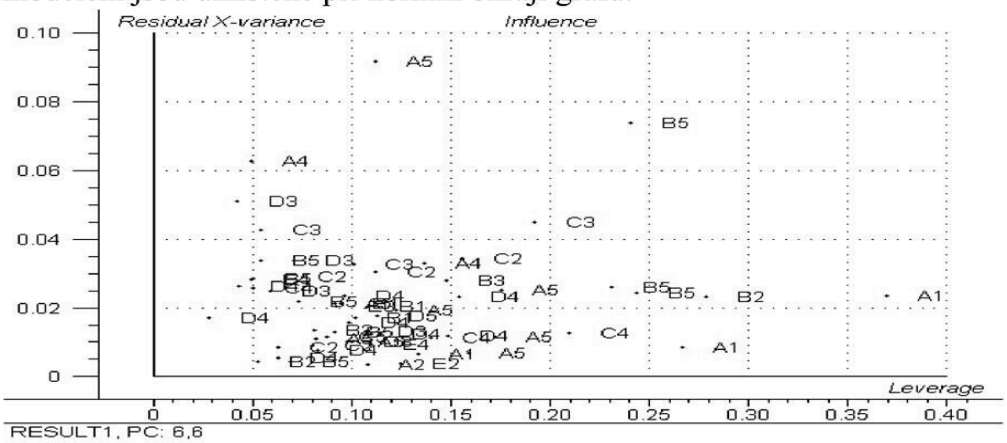
3. Rozptylový diagram komponentního skóre: Písmena *A, B, C, D* a *E* označují typ odrůdy hrachu, zatímco číslo 1, 2, 3, 4 a 5 značí čas sklizně. *PC*1 souvisí s časem sklizně.



Obr. 4.9 Rozptylový diagram komponentního skóre dat *Hrách*.



4. Graf vlivných bodů: objekty které nejsou dostatečně popsány PCA modelem jsou umístěné při horním okraji grafu.



Obr. 4.10 Graf vlivných bodů statistické analýzy reziduí objektů dat *Hrách*.

○ *Závěr:* byl posouzen graf komponentního skóre k roztřídění odrůd hrachu dle svých dvou dominantních vlastností, dle času sklizně a dle svých odrůd.

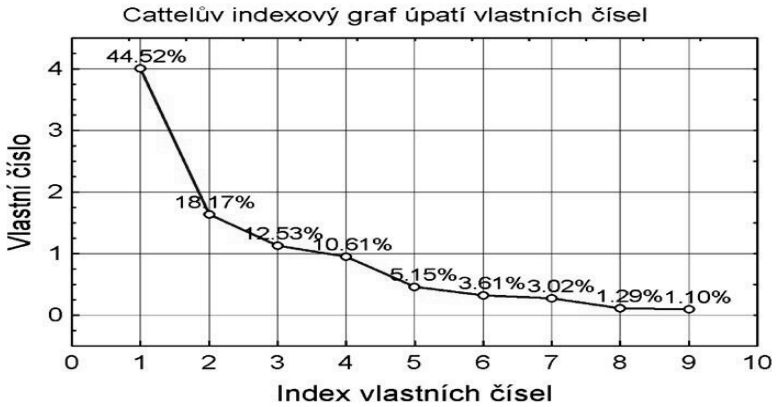
PŘÍKLAD 4.4 *Sledování spotřeby proteinů v zemích Evropy*

Sledována spotřeba proteinů v 25 zemích Evropy formou spotřeby 9 druhů potravin. Cílem je odhalit, zda existuje korelace mezi znaky, tj druhy potravin? Lze odhalit nějaké interakce mezi druhy potravin a zeměmi?

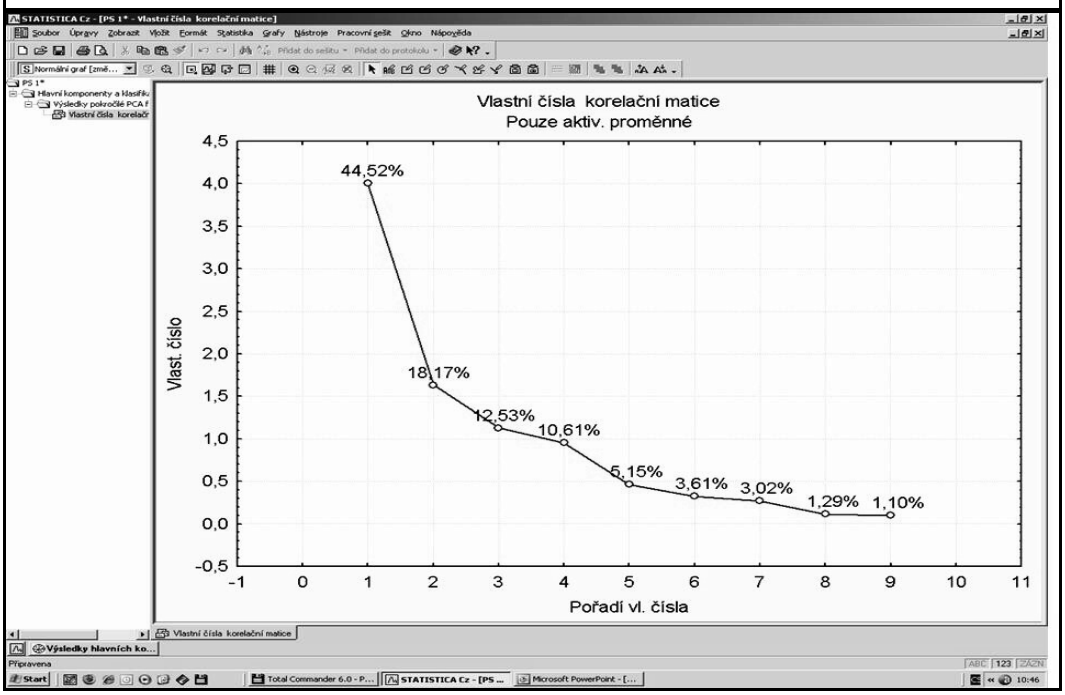
○ *Data:* v datech *Proteiny* jsou uvedeny znaky: *Cervene* značí spotřebu červeného masa, *Bile* značí spotřebu bílého masa, *Vejce* značí spotřebu vajec, *Mleko* se týká spotřeby mléka, *Ryby* značí spotřebu ryb, *Obiln* značí spotřebu obilnin, *Skrob* značí spotřebu škrobu, *Orech* značí spotřebu ořechů, *Ovoce* značí spotřebu ovoce a zeleniny.

Země	*Cervene*	*Bile*	*Vejce*	*Mleko*	*Ryby*	*Obiln*	*Skrob*	*Orech*	*Ovoce*
Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Austria	8.9	14	4.3	19.9	2.1	28	3.6	1.3	4.3
Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4
Bulgaria	7.8	6	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Czechoslovakia	9.7	11.4	2.8	12.5	2	34.3	5	1.1	4
Denmark	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4
East Germany	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Finland	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1	1.4
France	18	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
Greece	10.2	3	2.8	17.6	5.9	41.7	2.2	7.8	6.5
Hungary	5.3	12.4	2.9	9.7	0.3	40.1	4	5.4	4.2
Ireland	13.9	10	4.7	25.8	2.2	24	6.2	1.6	2.9
Italy	9	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Netherlands	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Norway	9.4	4.7	2.7	23.3	9.7	23	4.6	1.6	2.7
Poland	6.9	10.2	2.7	19.3	3	36.1	5.9	2	6.6
Portugal	6.2	3.7	1.1	4.9	14.2	27	5.9	4.7	7.9
Romania	6.2	6.3	1.5	11.1	1	49.6	3.1	5.3	2.8
Spain	7.1	3.4	3.1	8.6	7	29.2	5.7	5.9	7.2
Sweden	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2
Switzerland	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
UK	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
USSR	9.3	4.6	2.1	16.6	3	43.6	6.4	3.4	2.9
West Germany	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
Yugoslavia	4.4	5	1.2	9.5	0.6	55.9	3	5.7	3.2

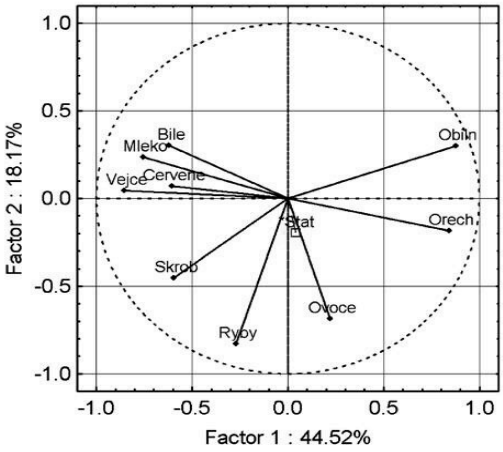
1. Cattelův indexový graf úpatí vlastních čísel: první hlavní komponenta (44.52% celkové proměnlivosti) a druhá hlavní komponenta (18.17% celkové proměnlivosti) dohromady dostatečně popíší proměnlivost v datech.



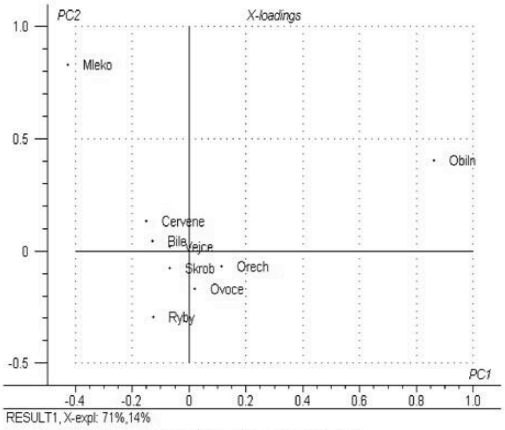
Obr. 4.15 Cattelův indexový graf úpatí celkového reziduálového rozptylu zdrojové matice dat *Proteiny* (STATISTICA).



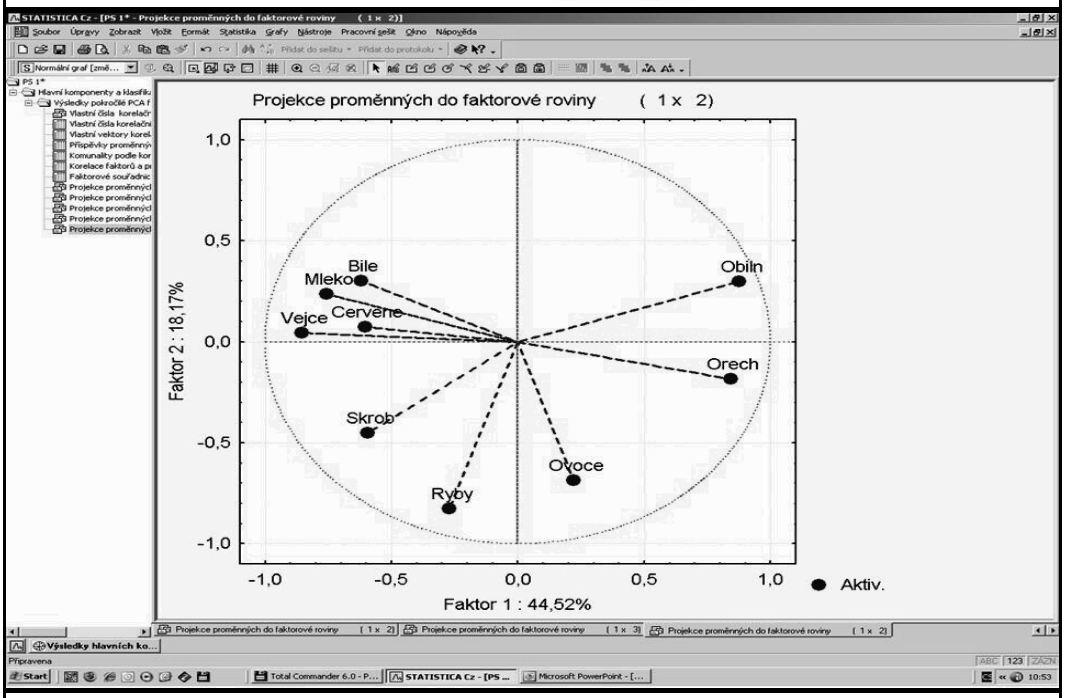
2. Graf komponentních vah: *Mléko* a *Obilniny* spolu vzhledem obsahu proteinů nekorelují. Vyjímečně si stojí i znak *Ryby*. Okolo počátku je shluk znaků, které jsou spolu v silné korelaci, jsou to *Červené* maso, *Bílé* maso, *Vejce*, *Škrob*, *Ořechy* a *Ovoce* a zelenina.



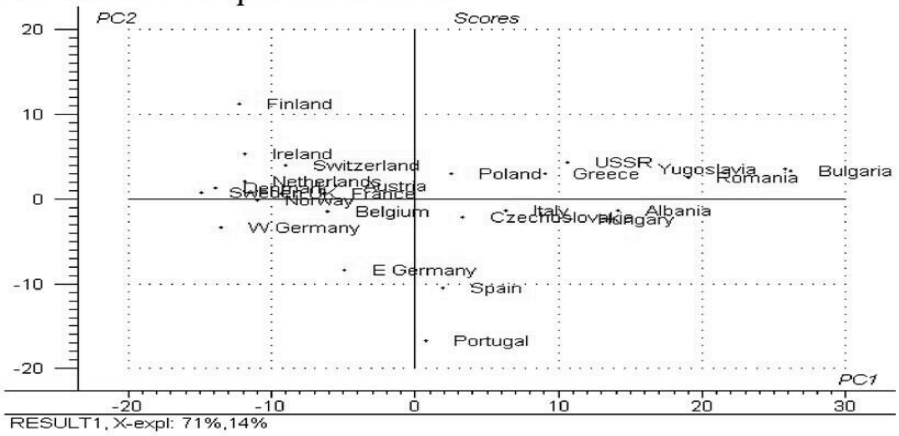
Obr. 4.16a Graf komponentních vah 1 a 2 dat *Proteiny* (STATISTICA).



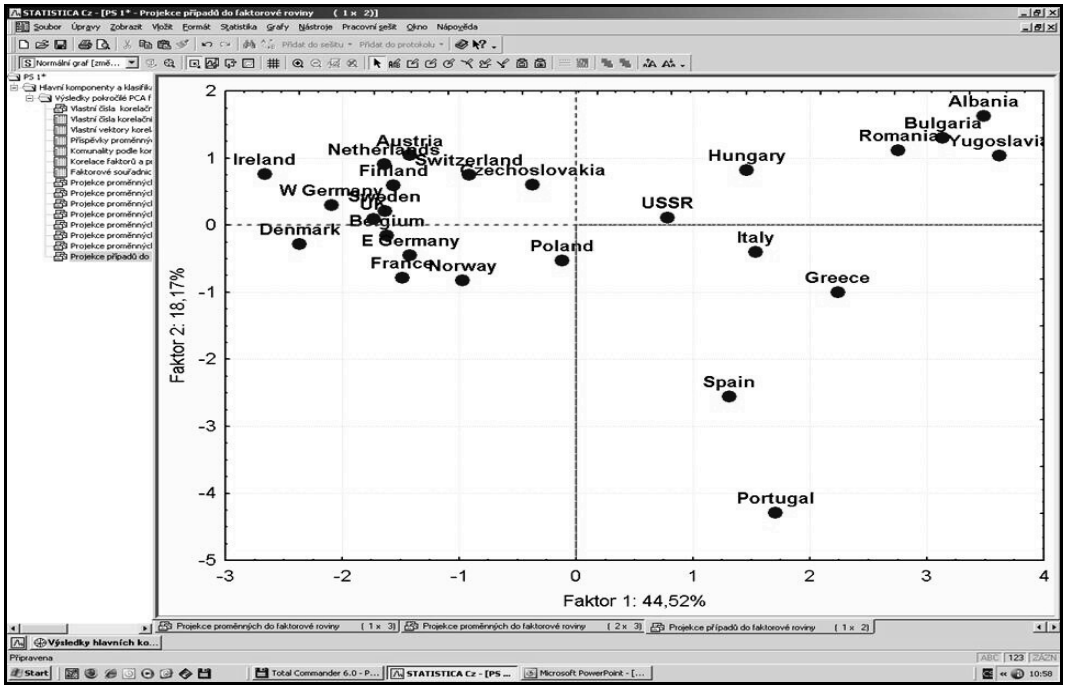
Obr. 4.16b Graf komponentních vah 1 a 2 dat *Proteiny* (UNSCRAMBLER).



3. Rozptylový diagram komponentního skóre: roztřídil státy dle spotřeby proteinů do shluků: shluk balkánských zemí (Bulharsko, Rumunsko, Albanie, Jugoslavie), shluk s zemí Polsko, Řecko, SSSR, Československo, Itálie a Maďarsko. Španělsko koreluje s Portugalskem a Východním Německem. Velký shluk obsahuje státy západní Evropy, ze kterých vybočuje Finsko a částečně i Západní Německo.



Obr. 4.17 Rozptylový diagram komponentního skóre dat *Proteiny* (UNSCRAMBLER).



Poděkování: Práce vznikla za podpory grantu Ministerstva zdravotnictví NR9055-4/2004 a vědeckých záměrů Ministerstva školství, kultury a mládeže MSMT0021627502.

Doporučená literatura

[1] Siotani M., Hayakawa T., Fujikoshi Y.: *Modern Multivariate Statistical Analysis*, A Graduate Course and Handbook. American Science Press, Columbia 1985.
[2] Chambers J. M., Cleveland W. S., Kleiner B., Tukey P. A.: *Graphical Methods for Data Analysis*. Duxburg Press, Belmont, California 1983.
[3] Barnett V., (Edit.): *Interpreting Multivariate Data*. Wiley, Chichester 1981, kap. 6.
[4] Jolliffe I. T.: *Principal Component Analysis*. Springer Verlag, New York 1986.
[5] Barnett V., (Edit.): *Interpreting Multivariate Data*. Wiley, Chichester 1981, kap. 12.
[6] Everitt B. S.: *Graphical Techniques for Multivariate Data*. London 1978.
[7] Johnson R.A., Wichern D.W.: *Applied Multivariate Statistical Analysis,* Prentice Hall, 1982
[9] Meloun M., Militký J. , Forina M.: *Chemometrics for Analytical Chemistry, Volume 1. PC-Aided Statistical Data Analysis,* Ellis Horwood, Chichester 1992.
[10] Brereton R. G. *Multivariate Pattern Recognition in Chemometrics, Illustrated by Case Studies*, Elsevier 1992,
[11] Krzanowski W. J.: *Principles of Multivariate Analysis, A User's Perspective*, Oxford Science Publications 1988,
[12] Meloun M. , Militký J., *Statistické zpracování experimentálních dat*, Plus Praha 1994, Academia Praha 2004..
[13] Meloun M. , Militký J., Hill M., *Počítačová analýza vícerozměrných dat v příkladech,* Academia Praha 2005.
[15] Meloun M. , Militký J., *Kompendium statistického zpracování experimentálních dat*, Academia Praha 2002, 2006.